J Am Med Inform Assoc 2020:(in press)

Assessing the Enrichment of Dietary Supplement Coverage in the UMLS

Jake Vasilakes[1,2], Anusha Bompelli[1], Jeffrey R. Bishop[3], Terrence J. Adam[1,2], Olivier Bodenreider[4], Rui Zhang[1,2]*

[1]Institute for Health Informatics, [2]Department of Pharmaceutical Care & Health Systems,

[3]Department of Experimental and Clinical Pharmacy, University of Minnesota, Minneapolis, MN, USA

[4] Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD, USA

Corresponding author:

*Dr. Rui Zhang, PhD

Department of Pharmaceutical Care & Health Systems and Institute for Health Informatics,

University of Minnesota

MMC 912, 420 Delaware St SE, Minneapolis, MN 5545

Email: zhan1386@umn.edu

Office Phone: 612-626-4209

Keywords:

Dietary supplements; Terminology; Unified Medical Language System, Named entity recognition

Word count: 4163

ABSTRACT

Objective: To assess the need for additional coverage of dietary supplements (DS) in the UMLS by investigating 1) the overlap between the iDISK DS ingredient terminology and the UMLS, and 2) the coverage of iDISK and the UMLS over DS mentions in the biomedical literature.

Materials and Methods: We estimated the overlap between iDISK and the UMLS by mapping iDISK to the UMLS using exact and normalized strings. The coverage of iDISK and the UMLS over DS mentions in the biomedical literature was evaluated via a DS named-entity recognition (NER) task within PubMed abstracts.

Results: The coverage analysis revealed that only 30% of iDISK terms can be matched to the UMLS, although these cover over 99% of iDISK concepts. A manual review revealed that a majority of the unmatched terms represented new synonyms, rather than lexical variants. For NER, iDISK nearly doubles the precision and achieves a higher F1 score than the UMLS, while maintaining a competitive recall.

Discussion: While iDISK has significant concept overlap with the UMLS, it contains many novel synonyms. Furthermore, almost 3000 of these overlapping UMLS concepts are missing a DS designation, which could be provided by iDISK. The NER experiments show that the specialization of iDISK is useful for identifying DS mentions.

Conclusion: Our results show that the DS representation in the UMLS could be enriched by adding DS designations to many concepts and by adding new synonyms.

BACKGROUND AND SIGNIFICANCE

The Dietary Supplement Health and Education Act (DSHEA) of 1994 defines dietary supplements (DS) as products ingested or administered to the body that contain a "dietary ingredient". This includes vitamins, minerals, amino acids, and herbs or botanicals, as well as other substances that can be used to supplement the diet [1]. 49% of the total U.S. population uses DS (males 44%, females 53%), according to the National Health and Nutrition Examination Survey (NHANES) [2]. DS are primarily considered as food, compared to prescription and over-the-counter drugs, and are regulated by the FDA under a different, less stringent set of rules.

The high utilization of DS alongside less stringent regulation compared to drugs, makes it difficult to navigate the terminology surrounding DS. Ingredients are referred to in a number of ways (e.g. *ginkgo, maidenhair tree, baiguo*), with derivative ingredients (e.g. *ginkgetin*) or product names (e.g. *Ultra! Ginkgo*) often used interchangeably. Additionally, high utilization of DS and relatively low regulation present challenges with respect to pharmacovigilance and drug safety monitoring. Drug-supplement interactions are one major risk, potentially putting DS users at risk of toxicity or therapeutic failure [3]. For example, there are over 30 known DS interactions with the anticoagulant warfarin [4] and St. John's Wort, an herbal DS commonly used for depression symptoms, may reduce the effectiveness of certain heart medications, antihypertensives, and lipid lowering agents [5]. As another example, one review noted a limited capacity to find DS mentions in primary resources supporting poison control center databases, hindering the ability to respond to emergency calls related to ingestions or possible poisoning events [6]. The ability to automatically extract DS mentions from information systems such as the biomedical literature and EHRs could improve patient outcomes by facilitating the curation of DS knowledge and flagging potential DS-related risks. To our knowledge, only one such system has been developed, supp.ai

[7]. supp.ai, however, does not contain a formal DS terminology and it is also closed-source, which limits its adoptability by the research community.

Standardized biomedical terminologies are important resources for facilitating cross-platform communicability and representing knowledge for healthcare providers, clinical researchers, and patients. The Unified Medical Language System (UMLS) [8] integrates a number of such terminologies, such as RxNorm [9], SNOMED CT [10], and MeSH [11] under a common framework, further improving interoperability. These terminologies are also often used for natural language processing (NLP) tasks such as named-entity recognition (NER) and information retrieval, providing seed terms to search for in the input text [12-14]. In the specific domain of DS, our prior studies have demonstrated limited coverage of DS terms in online resources [15], product labeling databases (e.g., DSLD) [16], and electronic health records [17]. Related work has provided use cases of enriching the UMLS on the specific domains. For example, a standard terminology for gastrointestinal endoscopy reporting, MST, was integrated in the 2002 UMLS Metathesaurus to support interoperability of clinical data repositories in gastroenterology [18]. Human Phenotype Ontology (HPO), mainly used for research purposes, was investigated to map to SNOMED CT, mainly used in healthcare, through lexical or logical mappings [19].

OBJECTIVE

The UMLS in its current state (version 2019AB) does not contain any resources specific to DS. Additionally, while the UMLS does cover many concepts that can be considered DS, such as vitamin C (C0003968) and ginkgo (C0330206), there is no ideal way of identifying these concepts as such. We therefore assess the need for enriching the coverage of DS in the UMLS. To do this, we employ the DS ingredient terminology contained in the integrated Dietary Supplements

Knowledge base (iDISK) [20]. Specifically, we analyze the coverage of iDISK against the UMLS and explore the applicability of using iDISK and UMLS terms for a DS named-entity recognition (NER) task on annotated PubMed abstracts.

MATERIALS AND METHODS

As illustrated in Figure 1, this study is broadly divided into two sets of experiments. First, we investigate the coverage of the iDISK DS ingredient terminology over both the entire UMLS and a subset of UMLS focused on DS. Second, in the NER experiments, we employ a keyword-lookup approach to evaluate the practical coverage of the iDISK and UMLS terminologies over mentions of DS ingredients manually annotated in 351 PubMed abstracts. We discuss the terminologies and each of these experiments in detail below.

Term Lists

This study employs four term lists, extracted from the UMLS and iDISK terminologies.

UMLS: We used the 2019AB version of the UMLS Metathesaurus.

$UMLS_{Distilled}$: We created a subset of the UMLS that contains only the following semantic types, which apply to the domain of DS: Amino Acid, Peptide, or Protein, Biologically Active Substance, Chemical, Element, Ion, or Isotope, Hormone, Immunologic Factor, Inorganic Chemical, Nucleic Acid, Nucleoside, or Nucleotide, Organic Chemical, Pharmacologic Substance, Vitamin, Bacterium, Fish, Fungus, Plant, or Food.

$UMLS_{DS}$: We further derived a subset of $UMLS_{Distilled}$ focused on dietary supplements, called $UMLS_{DS}$, by computing the transitive closure of hierarchical relationships in the Metathesaurus, using the technique suggested in [21]. Starting from the root concept, we add its

direct children (CHD) and narrower concepts (NAR). Then, recursively, we add their children and narrower concepts until no new concepts can be added. In this application, we use the "Dietary Supplements" (C0242295) and "Vitamin" (C0042890) concepts as the roots.

iDISK : We extracted the DS ingredient terminology contained in iDISK 1.0.1, which was built by automatically integrating four well-regarded DS resources: The Natural Medicines Comprehensive Database (NMCD), the "About Herbs" page on the Memorial Sloan Kettering Cancer Center (MSKCC) website, the Dietary Supplement Label Database (DSLD), and the Natural Health Products Database (NHP).

Task 1: Coverage analysis between iDISK and UMLS

The goal of these experiments is to investigate the alignment between the iDISK DS ingredient terminology and the UMLS. We extracted the term lists described above from iDISK version 1.0.1 and the UMLS Metathesaurus 2019AB by obtaining each term from each concept, converting it to lowercase, and removing any terms that fully match a stop word defined in the Python Natural Language Toolkit (NLTK) [22]. We also filtered terms that were a single letter, only numbers, or only punctuation. We used exact and normalized string matching to find which iDISK terms occur in the UMLS, $UMLS_{Distilled}$, and $UMLS_{DS}$ term lists. Normalized string matching was performed by first processing the iDISK terms with luiNorm from the SPECIALIST Lexical Tools[1] and searching for a match in the UMLS term list. luiNorm works by abstracting away word modifiers such as plural forms, case, punctuation, and word order. For example, luiNorm normalizes "Hedera helix leaves extract" to "extract hedera helix leaf".

[1] https://lsg2.nlm.nih.gov/Specialist/Home/index.html

In both the UMLS and iDISK, concepts are collections of atoms. An atom is a unique combination of a term, its source vocabulary (e.g. RxNorm), and its source ID. There may thus be multiple atoms with identical terms. As a consequence, we report the number and percentage of iDISK atoms and unique terms that were found in the UMLS, $UMLS_{Distilled}$, and $UMLS_{DS}$ term lists using exact and normalized string matching. We also explore the concept alignment between iDISK and the UMLS by computing the percentage of atoms matched per iDISK concept and the number of unique UMLS concepts to which they were matched.

Qualitative analysis of unmatched iDISK terms

We performed a qualitative analysis to determine whether the unmatched iDISK terms are lexical variants or novel synonyms of existing UMLS terms. An annotator with a PharmD and an informatics background (AB) performed the following steps for each term in a randomly sampled subset of 1000 iDISK terms unmatched to $UMLS_{Distilled}$.

1. Disambiguated the iDISK term to determine the best matching UMLS concept, if one did exist.
2. Compared the iDISK term with the atom terms of the UMLS concept to determine the variant type of the iDISK term. The possible variant types are
    a. Lexical: The iDISK term is a lexical variant of one or more atom strings of the UMLS concept.
    b. Synonym: The iDISK term is a novel synonym of the concept and no lexical variant of this term exists as an atom of the concept.

We report the number of terms per variant type and ingredient type.

Task 2: NER of DS mentions in PubMed abstracts

While the coverage experiments show how the UMLS and iDISK overlap, the NER experiments investigate how well each term list covers the variety of mentions present in biomedical text. The goal here is not to achieve the best performance on the NER task, but to explore the relative strengths and weaknesses of each term list in a practical setting. Thus, rather than using semantically and contextually aware NER methods such as neural sequence models or MetaMap (which uses a variety of heuristics), we used a simple keyword-lookup approach in order to isolate the effect of the term lists. We used an existing UMLS entity-linking tool, QuickUMLS [23], to perform the DS NER task as it does not use context, is designed for biomedical text, and is straightforward to modify. Unmodified QuickUMLS works by indexing the terms in the UMLS MRCONSO.RRF file to allow for quick look up by string similarity to the input text, outputting spans in the input that exceed a string similarity threshold to one or more UMLS terms. Our NER implementation thus involved creating custom keyword lists of DS, loading them into QuickUMLS, and running them on a set of PubMed abstracts, treating all predicted spans as DS mentions.

System Descriptions

We built the QuickUMLS NER installations using the UMLS, $\text{UMLS}_{\text{Distilled}}$, UMLS_{DS}, and iDISK term lists. We used the following QuickUMLS settings: the minimum Jaccard similarity between a term and a matched span was 0.7 out of 1.0. A matched span was allowed to have no more than 5 tokens. Overlapping matched spans were disambiguated according to which had the higher Jaccard similarity score. Additionally, we restricted QuickUMLS to only match spans with 3 or more characters. While this restriction would surely result in false negatives

(FNs), preliminary experiments showed that allowing spans with less than 3 characters resulted in many false positives (FPs). In addition to these NER systems, we also evaluate the union of the spans predicted by the individual $UMLS_{DS}$ and iDISK systems. This evaluation indicates the total number of true positives (TPs) possible using both DS-focused term lists and the types of terms not represented by either.

Annotation of PubMed abstracts

We obtained abstracts for the evaluation by querying PubMed for the MeSH term "dietary supplement" in the following academic journals: The Annual Review of Nutrition[2], Advances in Nutrition[3], the American Journal of Clinical Nutrition[4], Critical Reviews in Food Science and Nutrition[5], and Nutrition Reviews[6]. The exact query, submitted on January 1, 2019 is below.

```
(dietary supplement[MeSH Terms]) AND
("annual review of nutrition"[Journal] OR
  "advances in nutrition bethesda, md"[Journal] OR
  "the american journal of clinical nutrition"[Journal] OR
  "critical reviews in food science and nutrition"[Journal] OR
  "nutrition reviews"[Journal])
```

The results were filtered to those published within the previous 5 years, resulting in 482 abstracts. After removing duplicates and abstracts containing no DS mentions according to a manual review, 351 abstracts remained.

[2] https://www.annualreviews.org/journal/nutr
[3] https://academic.oup.com/advances
[4] https://academic.oup.com/ajcn
[5] https://www.tandfonline.com/toc/bfsn20/current
[6] https://academic.oup.com/nutritionreviews

The task of annotating the abstracts for DS mentions was split between two annotators with an MD (RR) and a PharmD (AB) degree. The initial annotation guidelines instructed the annotators to highlight a text span if it fits the definition of a DS according to the DSHEA, i.e. it falls under "vitamins, minerals, amino acids, and herbs or botanicals, as well as other substances that can be used to supplement the diet" [1]. These guidelines were then iteratively developed on a small subset of 10 abstracts to determine the best annotation for exceptional cases. Disagreements were ultimately resolved by a pharmacologist (JB). Once the guidelines were finalized, inter-annotator agreement was computed over 40 abstracts (~ 400 entity mentions), revealing a Cohen's kappa of 0.80 and a percentage agreement of 0.98, which indicate strong agreement [24]. The resulting gold standard contains 3710 annotations.

Evaluation

The QuickUMLS NER systems were evaluated against the gold standard annotations in the 351 PubMed abstracts described in the previous section. We report the precision, recall, and F1 score of each system over all the abstracts according to two evaluation criteria, adopted from the n2c2 2018 Track 2 entity extraction challenge [25]:

- Lenient: A predicted span is a TP if it overlaps the gold standard by any number of characters.
- Strict: A predicted span is a TP if and only if it exactly matches a gold standard span.

To further compare system performance, we report the number of TPs, FPs, and FNs according to both evaluation criteria, as well as the total number of predicted spans. We also performed an error analysis by manually reviewing the FPs and FNs.

RESULTS

Coverage Experiments

The full UMLS term list contains 12 268 715 unique terms corresponding to 4 258 810 concepts. $UMLS_{Distilled}$ contains 2 738 437 unique terms corresponding to 1 275 143 concepts, and is about 30% of the full UMLS. $UMLS_{DS}$ contains 1 182 552 unique terms corresponding to 345 376 concepts, about 8% of the full UMLS. The iDISK term list contains 41 628 unique DS ingredient terms corresponding to 4101 concepts.

Table 1 provides the number of iDISK atoms and unique terms that were matched to each of the UMLS term lists using exact matching as well as the number of matches gained from string normalization. We also report the total number of UMLS concepts matched per term list. It shows that iDISK matches a similar number of atoms and terms in the UMLS and $UMLS_{Distilled}$ term lists, even though $UMLS_{Distilled}$ contains only around 30% of the UMLS concepts. However, an additional 2000 UMLS concepts were matched in the UMLS term list over $UMLS_{Distilled}$. A preliminary review suggests that many DS terms in iDISK are ambiguous when matched against the full UMLS. For example, “Aki” is correctly matched to the DS concept “Blighia sapida” (C0330978), but also to the non-DS concepts “Acute Kidney Injury” (C2609414) and “Kidney Failure, Acute” (C0022660), which include the acronym “AKI” among their atoms.

Table 1: The number and percentage of iDISK atoms and unique terms matched against the UMLS, $UMLS_{Distilled}$, and $UMLS_{DS}$ term lists using both exact and normalized matching with luiNorm. The iDISK term list contains 61 228 atoms, corresponding to 41 628 unique DS ingredient terms. The percentages in the “Atoms” and “Unique Terms” rows are out of these totals,

respectively. The numbers in the "UMLS Concepts" column indicate the number of unique UMLS concepts that correspond to the total number of matched atoms.

Matched Against	iDISK element	Exact Match (%)	+luiNorm (+%)	Total (%)	UMLS Concepts
UMLS	Atoms	27 992 (45.7%)	+550 (+0.9%)	28 542 (46.6%)	10 716
	Unique Terms	12 744 (30.6%)	+474 (+1.1%)	13 218 (31.7%)	
$UMLS_{Distilled}$	Atoms	27 553 (45.0%)	+524 (+0.9%)	28 077 (45.9%)	8 684
	Unique Terms	12 397 (29.8%)	+450 (+1.0%)	12 847 (30.8%)	
$UMLS_{DS}$	Atoms	12 096 (19.7%)	+407 (+0.7%)	12 503 (20.4%)	2 813
	Unique Terms	4 899 (11.8%)	+308 (+0.7%)	5 207 (12.5%)	

We further explored the concept alignment between iDISK and the UMLS term lists by calculating (1) the percentage of iDISK atoms per concept mapped to each UMLS term list (Figure 2) and (2) number of UMLS concepts in each term list matched per iDISK concept (Figure 3). Figure 2 shows that nearly all iDISK concepts have some, but not all, of their atoms mapped to the full UMLS and $UMLS_{Distilled}$. However, half of the iDISK concepts (2079) are entirely unmatched to $UMLS_{DS}$. The heavy tailed distributions in Figure 3 show that many iDISK concepts are matched to multiple UMLS concepts.

Qualitative analysis of unmatched iDISK terms

Figure 2 shows that 839 out of 4101 iDISK concepts had 100% of their atoms matched to the full UMLS. This suggests that the remaining 3258 iDISK concepts, which had between 1% and 99% of their atoms matched, contain novel lexical variants and synonyms of existing UMLS concepts. To investigate whether this is indeed the case, we performed a qualitative analysis to

determine the type of terms missing in the UMLS but present in iDISK. As described in the Materials and Methods section, an annotator (AB) manually reviewed a random subset of 1000 iDISK terms that were unmatched to the full UMLS, disambiguated each term to determine the best matching UMLS concept, and compared the iDISK term to the UMLS atoms of that concept to determine whether the iDISK term was an unmatched lexical variant of an existing UMLS term or an entirely novel synonym. Using the UMLS semantic types of the disambiguated UMLS concepts, we determined their corresponding UMLS semantic group and compared these groups to the annotations. The results of this analysis are given in Table 2. In some cases, such as ambiguous acronyms or non-specific plant parts, it was not possible to disambiguate a term. These cases were marked as "Ambiguous" and no variant type was annotated.

Table 2: The distributions of variant type ("Lexical variant" or "Synonym") and semantic group ("Chemicals and Drugs" or "Living Beings") for 1000 manually reviewed iDISK terms that were unmatched to $UMLS_{All}$. Percentages are out of the total number of terms described in that column.

	Lexical variant	Synonym	Ambiguous
Chemicals and Drugs	132 (55%)	137 (19%)	-
Living Beings	109 (45%)	599 (81%)	-
Total	241	736	23

Table 2 shows that a majority (736, around 73%) of the unmatched iDISK terms are novel synonyms, rather than lexical variants, of UMLS concepts. A deeper review of the novel synonyms showed that they are often common or synonymous binomial names of plants. For instance, the UMLS concept C3927248 contains only the binomial name "Calea ternifolia" and no other synonyms, yet the manual review mapped approximately 50 unmatched iDISK terms to this

concept, such as “bitter-grass” and “Calea rugosa”. The terms belonging to the “Chemicals and Drugs” semantic group were primarily different forms of chemical or molecular nomenclature.

As shown in Figure 2, only 4 iDISK concepts were completely unmatched to the UMLS. 2 are indeed present in the UMLS, but there were no overlapping synonyms. For example, the UMLS contains the term “Paris polyphylla” (C1014802), but iDISK only contains “Paris polyphylla var. yunnanensis”. For 1 iDISK concept, “Hedera helix extract”, only more specific concepts are present in the UMLS, i.e. “Hedera helix leaf extract” (C3255599) and “Hedera helix flowering twig extract” (C3486554). Finally, 1 iDISK concept, “Wild Berry”, was too unspecific to be matched to any UMLS concept.

A number of $UMLS_{DS}$ concepts (339 559) were not matched to iDISK, potentially indicating DS concepts that iDISK does not cover. We manually reviewed 400 randomly sampled concepts (stratified by semantic types) from this unmatched set and identified that 152 (38%) were DS-related. Among these 152 concepts, only 15 were, in fact, DS ingredients that are missing from iDISK. Eighty were brand names of DS products and were not included in this analysis, which focused on ingredients (although iDISK does contain DS product names). Thirty-three had a corresponding iDISK concept, but no match was possible using our methods. Finally, 24 were general DS descriptions or blends, such as “Amino Acid/Electrolyte Mixture-based Dietary Supplement” (C4329299) and had no corresponding iDISK entry.

Named-Entity Recognition Experiments

Table 3 reports the precision, recall, and F1 score for each NER system computed over all 351 PubMed abstracts. Figure 4 gives details on the distribution of these metrics across individual

abstracts. In order to further evaluate the NER systems, we compare the number and type of predictions made by each system over all the abstracts in Table 4.

Table 3: Precision, recall, and F1 score for the 5 QuickUMLS NER systems evaluated against the gold standard annotations from 351 PubMed abstracts. The best results for each metric are given in bold. *Union: the set union of the predictions made by the $UMLS_{DS}$ and iDISK systems.

Evaluation Criterion	QuickUMLS System	Precision	Recall	F1
Lenient	UMLS	0.08	**0.91**	0.15
	$UMLS_{Distilled}$	0.25	0.89	0.39
	$UMLS_{DS}$	0.32	0.86	0.46
	iDISK	**0.51**	0.82	**0.63**
	Union*	0.32	**0.91**	0.48
Strict	UMLS	0.05	0.67	0.10
	$UMLS_{Distilled}$	0.19	0.69	0.30
	$UMLS_{DS}$	0.22	0.61	0.33
	iDISK	**0.43**	0.69	**0.53**
	Union*	0.23	**0.77**	0.36

The iDISK system achieved much higher precision (0.51 lenient and 0.43 strict) than any other system, and nearly doubled the precision of the next best system ($UMLS_{DS}$ at 0.22) under the strict evaluation criterion. This increase in precision was due to many fewer false positives than the other systems (3427 by iDISK vs. 7815 by $UMLS_{DS}$), suggesting that the iDISK term list is more specialized than the others. Despite its specialization and significantly smaller vocabulary (0.3% the size of the full UMLS), iDISK obtained a competitive recall of 0.69 under the strict

criterion, outperforming the full UMLS and tying with $UMLS_{Distilled}$. It also achieved 0.82 recall under the lenient criterion.

Table 4: The number of True Positive (TP), False Positive (FP), and False Negative (FN) predictions as well as the total number of predictions for each of the 4 QuickUMLS NER systems evaluated against the 351 gold standard PubMed abstracts. The best result for each prediction type is given in bold. For TPs higher is better. For FPs and FNs lower is better. *Union: the set union of the predictions made by the $UMLS_{DS}$ and iDISK systems.

Evaluation Criterion	QuickUMLS System	TP	FP	FN	Total Predictions (=TP+FP)
Lenient	UMLS	3577	41 795	**333**	45 372
	$UMLS_{Distilled}$	3430	10 167	405	13 597
	$UMLS_{DS}$	3180	6881	516	10 061
	iDISK	3044	**2941**	671	5 985
	Union*	**3982**	8308	413	12 290
Strict	UMLS	2495	42 877	1215	45 372
	$UMLS_{Distilled}$	2577	11 020	1133	13 597
	$UMLS_{DS}$	2246	7815	1464	10 061
	iDISK	2558	**3427**	1152	5985
	Union*	**2856**	9434	**854**	12 290

The Union system, which takes predictions from both the iDISK and $UMLS_{DS}$ systems, failed to outperform either individual system in terms of F1 score. Table 4 shows that this was due to a large increase in false positives (FPs) over the individual systems (e.g. 9434 strict FPs by the Union system vs. 7815 and 3427 FPs by the $UMLS_{DS}$ and iDISK systems, respectively). This

suggests that the $UMLS_{DS}$ and iDISK systems make different types of errors, which is expected given that only 11.8% of the unique iDISK terms were found in the $UMLS_{DS}$ term list, according to the coverage experiments. The evaluation of the Union system also indicates the upper bound of the true positives (TPs) using both DS-focused term lists. Table 4 shows that the number of TPs increases dramatically in the Union system according to both evaluation criteria. These results suggest that iDISK alone can accurately find DS terms (i.e. high precision), while the union of iDISK and the UMLS could retrieve more DS terms (i.e. high recall).

DISCUSSION

The UMLS has been used to create more effective and interoperable biomedical information systems. However, the current version of UMLS does not have a clear way (e.g., via semantic types or relationships) to identify DS concepts, making it challenging for consumers, clinicians, and researchers to further search and explore DS related knowledge, such as usage [26], adverse events [27], and drug interactions [7]. The difference in matched terms between iDISK and $UMLS_{DS}$ revealed 2867 concepts in the UMLS which are potentially missing a DS designation. These concepts include “Centaurium erythraea” (C3472714, used for GI discomfort and fever), “Barley” (C0450033, used for high cholesterol), and “Phellodendron amurense bark extract” (C3256251, used in traditional Chinese medicine as an anti-inflammatory and antioxidant). Additionally, 99% of iDISK concepts had one or more term matched to the UMLS, and Figure 3 shows that many are matched to more than 1 UMLS concept, which suggest that iDISK and the UMLS represent DS concepts differently. Still, over 70% of the iDISK terms were not matched. A qualitative analysis of these unmatched terms revealed that iDISK covers a variety of DS

concepts and their synonyms, especially common names, which are otherwise not present in the UMLS.

NER Error Analysis

We performed a qualitative error analysis by manually reviewing the following incorrect spans and categorizing the reasons they were incorrect, plotting the results in Figure 5: (1) All 413 FN spans missed by the Union system under the lenient evaluation, as these represent terms that are not covered by either $UMLS_{DS}$ or iDISK (Figure 5-A), and (2) 424 out of 2941 (14%) FP spans generated by the iDISK system, as it is the most specialized of the term lists in terms of precision (Figure 5-B).

As shown in Figure 5-A, a majority (72%) of the FNs were simply not present in either term list. Additionally, 20% of the FNs are due to our requirement that matched spans be at least 3 characters, meaning mentions such as “Fe” for iron are missed. A small number (8%) of FNs were due to tokenization issues. For example, “EPA” and “DHA” were annotated as distinct mentions in the gold standard for “EPA+DHA”, but QuickUMLS did not split on the “+”, treating the input as a single token which did not match any terms in either list.

Figure 5-B shows that 39% of the FPs reviewed covered terms that are not, in fact, DS, such as “diet” and “salt”. A small number are DS terms, but were too generic to be annotated in the gold standard, such as “enzymes”. Nearly half of the FPs should not have been predicted due to context. A majority of these contexts were about serum concentration (e.g. measurement, deficiency) or a description of a biochemical process (e.g. metabolism).

Limitations and Future Work

One limitation of this work is the creation of the $UMLS_{DS}$ term list, which computes the transitive closure of the Dietary Supplement and Vitamin concepts following a similar work [19 21]. However, the transitive closure produced a number of non-DS concepts. For example, *Insulin* is incorrectly included in this list via the following path: *Dietary Supplement* (C0242295) → *Amino Acid Supplement* (C0556082) → *Amino Acids* (C0002520) → *Proteins* (C0033684) → *Peptide Hormone* (C0597192) → *Insulin* (C0021641). Future work could be to manually filter non-DS concepts from this list, as was done in [7], to obtain a more accurate comparison between iDISK and the UMLS. Another limitation is that we only used exact and normalized matching to align iDISK to the UMLS. This may limit our estimate of the coverage of DS terms in UMLS. In the future, we may extend the coverage analysis by adapting existing methods, including partial matching [19 28 29] and embedding methods [30].

The goal of the NER experiments was to isolate and evaluate the effective differences between the term lists for the identification of DS mentions, rather than to evaluate QuickUMLS as a practical solution for DS NER. While QuickUMLS provided a fast and interpretable framework for carrying out these experiments, our error analysis showed that its limited ability to account for context produced a number of false positive predictions. The NER results given here are thus lower than what could likely be achieved using state-of-the-art machine learning methods, such as BioBERT [31]. It is important to note, however, that such methods would require a significant amount of annotated training data, where QuickUMLS requires none. Once such a training dataset is compiled, future work would be to compare the effects of these different term lists on this NER task using these more robust machine learning methods. Finally, we did not characterize coverage relationships between the UMLS and iDISK as it was out of the

scope of this work, which focused on term and concept coverage. However, relationships are an integral part of both the UMLS and iDISK, so such an investigation is important future work.

CONCLUSION

The UMLS currently contains no resource dedicated to the growing domain of DS. This work assessed the need for enriching the UMLS with additional DS terms and concepts, using iDISK, which contains a terminology of DS ingredient terms. We found that although concept overlap between iDISK and the UMLS is high, iDISK contains many additional DS terms, and that most of these missing terms are entirely novel synonyms of UMLS concepts rather than lexical variants of existing terms. Furthermore, the iDISK terminology outperformed (in precision and F1 score) the UMLS Metathesaurus when used for a DS NER task, demonstrating the value of a more specialized DS representation for information extraction and retrieval tasks in the DS domain. Together, these results demonstrate the value of enriching the DS coverage in the UMLS.

DATA AVAILABILITY

The code, data, and annotation guidelines used in this study are made publicly available at https://github.com/zhang-informatics/UMLS_iDISK.

ACKNOWLEDGEMENTS

We would like to thank Dr. Rubina Rizvi for her annotation of the PubMed abstracts used in the NER evaluation and Nan Wang for his evaluation of $UMLS_{DS}$ concepts.

FUNDING STATEMENT

This work was supported by the National Institutes of Health's National Center for Complementary & Integrative Health (NCCIH) and the Office of Dietary Supplements (ODS) grant number R01AT009457 (Zhang), the National Center for Advancing Translational Sciences grant number UL1TR002494, and the Intramural Research Program of the National Library of Medicine. The content is solely the responsibility of the authors and does not represent the official views of the NIH.

COMPETING INTERESTS STATEMENT

The authors state that they have no competing interests to declare.

CONTRIBUTORSHIP STATEMENT

JV and RZ conceived the study design and wrote much of the manuscript. JV carried out the coverage and NER experiments and produced the original draft of the manuscript. AB annotated the PubMed abstracts and conducted the qualitative analysis of the coverage experiments. OB consulted on specifics of the experimental design. All authors contributed to the production of the manuscript.

REFERENCES

1. Dietary Supplement Health and Education Act of 1994, 1994:Pub L. No. 103--417.
2. Bailey RL, Gahche JJ, Lentino CV, et al. Dietary Supplement Use in the United States, 2003–2006. The Journal of Nutrition 2011;**141**(2):261 doi: 10.3945/JN.110.133025.
3. Lee AH, Ingraham SE, Kopp M, Foraida MI, Jazieh AR. The incidence of potential interactions between dietary supplements and prescription medications in cancer patients at a Veterans Administration Hospital. Am J Clin Oncol 2006;**29**(2):178-82 doi: 10.1097/01.coc.0000209369.44100.25.
4. Nutescu E, Chuatrisorn I, Hellenbart E. Drug and dietary interactions of warfarin and novel oral anticoagulants: an update. J Thromb Thrombolysis 2011;**31**(3):326-43 doi: 10.1007/s11239-011-0561-1.
5. Tsai HH, Lin HW, Simon Pickard A, Tsai HY, Mahady GB. Evaluation of documented drug interactions and contraindications associated with herbs and dietary supplements: a systematic literature review. Int J Clin Pract 2012;**66**(11):1056-78 doi: 10.1111/j.1742-1241.2012.03008.x.
6. Palmer ME, Haller C, McKinney PE, et al. Adverse events associated with dietary supplements: an observational study. Lancet 2003;**361**(9352):101-6 doi: 10.1016/S0140-6736(03)12227-1.
7. Wang LL, Tafjord O, Jain S, et al. Extracting evidence of supplement-drug interactions from literature. arXiv preprint 2019;arXiv:1909.08135.
8. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic acids research 2004;**32**(Database issue):D267--70 doi: 10.1093/nar/gkh061.
9. RxNorm Technical Documentation. 2018 01/02/2018; 2018. https://www.nlm.nih.gov/research/umls/rxnorm/docs/2018/rxnorm_doco_full_2018-1.html (accessed 2020).
10. SNOMED CT. Secondary SNOMED CT. www.snomed.org (accessed 2020).
11. Medical Subject Headings (MeSH). Secondary Medical Subject Headings (MeSH). meshb.nlm.nih.gov.
12. Kim S, Lu Z, Wilbur WJ. Identifying named entities from PubMed for enriching semantic categories. BMC Bioinformatics 2015;**16**:57 doi: 10.1186/s12859-015-0487-2.
13. Tran LT, Divita G, Carter ME, Judd J, Samore MH, Gundlapalli AV. Exploiting the UMLS Metathesaurus for extracting and categorizing concepts representing signs and symptoms to anatomically related organ systems. J Biomed Inform 2015;**58**:19-27 doi: 10.1016/j.jbi.2015.08.024.
14. Srinivasan S, Rindflesch TC, Hole WT, Aronson AR, Mork JG. Finding UMLS Metathesaurus concepts in MEDLINE. Proc AMIA Symp 2002:727-31.
15. Manohar N, Adam TJ, Pakhomov SV, Melton GB, Zhang R. Evaluation of Herbal and Dietary Supplement Resource Term Coverage. Studies in health technology and informatics 2015;**216**:785--9.
16. Wang Y, Adam TJ, Zhang R. Term Coverage of Dietary Supplements Ingredients in Product Labels. AMIA Annual Symposium 2016;**2016**:2053--61.

17. Zhang R, Manohar N, Arsoniadis E, et al. Evaluating Term Coverage of Herbal and Dietary Supplements in Electronic Health Records. Proceedings of the American Medical Informatics Association Symposium. 2015:1261-370.
18. Tringali M, Hole WT, Srinivasan S. Integration of a standard gastrointestinal endoscopy terminology in the UMLS Metathesaurus. Proc AMIA Symp 2002:801-5.
19. Dhombres F, Bodenreider O. Interoperability between phenotypes in research and healthcare terminologies-Investigating partial mappings between HPO and SNOMED CT. Journal of Biomedical Semantics 2016;**7**(1) doi: 10.1186/s13326-016-0047-3.
20. Rizvi RF, Vasilakes J, Adam TJ, et al. iDISK: the integrated DIetary Supplements Knowledge base. J Am Med Inform Assoc 2020;**27**(4):539-48 doi: 10.1093/jamia/ocz216.
21. Bodenreider O, Nelson SJ, Hole WT, Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. Proc AMIA Symp 1998:815-9.
22. Loper E, Bird S. NLTK: The Natural Language Toolkit. Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002:63--70.
23. Soldaini L, Goharian N. QuickUMLS: a fast, unsupervised approach for medical concept extraction. MedIR Workshop, Special Interest Group on Information Retrieval (SIGIR), 2016.
24. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;**22**(3):276-82.
25. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. Journal of the American Medical Informatics Association 2020;**27**(1):3--12 doi: 10.1093/jamia/ocz166.
26. Fan Y, Zhang R. Using natural language processing methods to classify use status of dietary supplements in clinical notes. BMC Med Inform Decis Mak 2018;**18**(Suppl 2):51 doi: 10.1186/s12911-018-0626-6.
27. Vasilakes JA, Rizvi RF, Zhang J, Adam TJ, Zhang R. Detecting Signals of Dietary Supplement Adverse Events from the CFSAN Adverse Event Reporting System (CAERS). AMIA Jt Summits Transl Sci Proc 2019;**2019**:258-66.
28. Huang KC, Geller J, Halper M, Cimino JJ. Piecewise synonyms for enhanced UMLS source terminology integration. AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium 2007:339--43.
29. Rudniy A, Geller J, Song M. Shortest Path Edit Distance for Enhancing UMLS Integration and Audit. AMIA Annual Symposium proceedings. AMIA Symposium 2010;**2010**:697--701.
30. Kolyvakis P, Kalousis A, Smith B, Kiritsis D. Biomedical ontology alignment: an approach based on representation learning. J Biomed Semantics 2018;**9**(1):21 doi: 10.1186/s13326-018-0187-8.
31. Lee J, Yoon W, Kim S, Kim D, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;**36**(4):1234-40 doi: 10.1093/bioinformatics/btz682.

FIGURE LEGENDS

Figure 1: Overview of this study.

Figure 2: Distribution of the percentage of atoms across iDISK concepts matched to each UMLS term list.

Figure 3: Distribution of the number of UMLS concepts matched in each UMLS term list across iDISK concepts with 1 or more atoms matched. The Y axis is log scaled for clarity.

Figure 4: Distributions of precision, recall, and F1 score according to the lenient and strict evaluation criteria on the 351 PubMed abstracts for each NER system.

Figure 5: Categorization of errors made by the NER systems. A) The categorization of all 413 false negative errors made by the Union (iDISK + $UMLS_{DS}$) system. B) The categorization of 424 out of 2941 (14%) false positive errors made by the iDISK system. We further subcategorize the errors due to context.